Adaptive Piecewise-constant Modeling of Signals in Multidimensional Spaces

J. Scargle
NASA Ames Research Center, Moffett Field, CA 94035, USA
B. Jackson
Department of Mathematics, San Jose State University; Center for Applied Mathematics and Computer Science
J. Norris
NASA Goddard Space Flight Center, Greenbelt, MD, 20771, USA

This contribution describes an analysis method for the class of problems in which data elements – *e.g.* measurements, event detections, *etc.* – are distributed over some region of space and/or time, or other coordinates (*e.g.*, energy, redshift, category), with the goal of estimating the variation of some physical quantity. The nonparametric model is simply that the physical variable is constant over a finite set of segments of the data space. A dynamic programming algorithm implements such modeling of 1D data by yielding the optimal partition of an interval. Any fitness function that is additive on the partition elements can be used, but the Bayesian posterior probability distribution over partitions—marginalized over all but the geometrical parameters defining the partition—has proved particularly effective. The resulting *maximum a posteriori* piecewise constant model is readily extended to data spaces of higher dimension.

1. Signal and Density Estimation

A goal of most astronomical observations and particle physics experiments is to describe the variation of some physical quantity as a function of time, space, energy, or other independent variable. We call such a function a *signal*. The experimental procedure is to measure the quantity at a finite number of points in the corresponding *data space*. This paper outlines a way to characterize signal variability using a simple nonparametric model of such data.

Related work on cluster detection in point data in the form of 2D catalogs can be found in [5, 8, 11]. These authors use the distribution of areas of Voronoi cells to establish a threshold of cell density below which lies background and above which are over densities, or clusters.

In subsequent sections we discuss the data (defining the key concept of *data cells*), segmented models and the corresponding fitness functions, the prior distribution for the number of blocks, the algorithm implementing the optimization, and finally application of the method to the large scale distribution of galaxies.

2. Data Cells

The data may be in any of a number of forms, such as points (*e.g.* galaxy positions), counts (*e.g.* particle events), measurements (*e.g.* spectral energy density), *etc*, as long as the measurement errors are independent. The current formalism cannot deal with dependent errors nor deconvolve the effects of dispersion—e.g., the *point spread function* affecting GLAST photon data.

The data points are to be associated with *data cells*. A simple example of a data cell is a bin—commonly used to estimate the density of points on some measurement axis. The complete description of the cell corresponding to a given bin requires specification of the number of samples in the bin, plus the bin's boundaries. More generally a data cell is a data structure representing an individual measurement within the *data space*—the set of all values that the measured quantity can possibly take on. For our segmented models, the cells must contain whatever information is needed to compute the model fitness function (§4).

In most cases it is natural to define the data cells to be in one-to-one correspondence to the measurements. But in a specific application it may be preferable to do otherwise—for example, if two or more events have the same time-tag, it may be reasonable to assign them to the same data cell. Similarly, in most cases it will be natural that the data cells partition the entire data space, with no overlap or gaps between cells; and typically the data cells contain information on adjacency to other cells. But in specific applications any of these conditions may be violated.

3. Piecewise Constant Models

We consider only segmented, *piecewise constant* models. That is to say the data space, whatever its dimension, is partitioned into a finite number of *blocks* within which the measured variable is represented as constant. The complete model consists of the partition, specified by the number of blocks N_b, a list of data cells in each blocks, plus the corresponding levels (*e.g.* event rates) in the blocks.

Boundaries separating blocks can be arbitrary: in 1D, points anywhere in the interval; in 2D, arbitrary line segments; in dimension ν, arbitrary surfaces of

dimension $\nu - 1$. Optimization over all possible partitions then involves hugely infinite search spaces.

However, a simple restriction on the class of allowed boundaries yields finite search spaces that are good approximations to the true ones, and turns the problem into a comparatively simple combinatorial optimization. The underlying idea is that two partitions differing only in a small distortion of a block boundary are not significantly different from each other. Construct a bounded volume element around each of the N data points, say consisting of that part of the data space closer to the point than to any other. In only a slight abuse of terminology, we associate these volumes with the data cells discussed above. Further, *blocks* are defined as sets of these cells. Correspondingly a partition of the whole data space is defined by collecting the N cells into distinct blocks. The set of all possible such assignments is finite, but represents an approximation to the hugely infinite set of all possible partitions.

If the cells are defined as above, they form what is called the *Voronoi tessellation*, a geometric partition easily computed in spaces of any dimension [10]. The partition elements, here called *blocks*, are simply sets of cells, with the condition that each cells belongs to one block, and not more than one. There are two important cases: the cells in a block must all be adjacent to each other, or one may not insist on this condition. Think of the blocks as analogous to level surfaces for an unknown function; the two cases correspond to distinguishing or identifying the disconnected parts of a given level surface.

Such step functions comprise the simplest class of nonparametric models, are very easy to interpret, and allow easy computation of summary physical quantities. In visualizations the choppiness due to discontinuities in the *block representation* can be ameliorated, *e.g.* by smoothing, if desired.

We want the model to be sensitive to any and all true variations, but insensitive to apparent variations produced by the inevitable observational errors.[1] We would like to preserve all features in the signal, on all scales supported by the data. But of course all analysis schemes—even those using nonparametric models—involve choices which restrict the questions that can be addressed. Our approach favors local structures over global ones. Because we want to be sensitive to features on fine as well as coarse scales, we do not use smoothing for noise suppression, but rather rely on the accuracy of the statistical model of the observational noise to effect **denoising without smoothing**.[2] A subsidiary goal is to implement an objective procedure suitable for automatic analysis of large data sets (data mining) such as those generated by modern particle physics and astrophysics projects.

The setting just described is more general that it perhaps first appears, and the methodology given here applies to a variety of seemingly different problems, and with a variety of distinct data types. The former include detection of signals and upper limits thereof, density estimation (usually for point data), detection and characterization of clusters, unsupervised classification, and others – including multivariate versions of any of these problems. Essentially any data mode can be treated, as long as one can compute a suitable fitness function for the block model. Fitness functions for point, binned count, and measurement data are readily computed, and categorical data can certainly be dealt with too. Distortions such as data gaps, variable instrumental sensitivity, and (at least in 1D) convolution with an instrumental point-spread function, can also be treated in very natural ways. Perhaps most useful of all is the ready treatment of data in any dimension.

4. Fitness Functions: Posterior Probabilities

A key element in implementing the modeling procedure is a function to measure goodness-of-fit for partitions. The standard Bayesian model estimation method yields convenient expressions valid for a variety of data modes. The simplicity of the block model makes such computations very easy. In particular, we need only compute the posterior for a single block, since statistical independence of the observational errors insures that the posterior for the whole data space is the product of that for each of the partition elements. Indeed, our algorithm requires additivity: the fitness of a partition must be the sum of the fitnesses of its blocks. This condition is achieved by using logarithms of posteriors.

Here is an outline of the procedure. The full posterior probability for the piecewise constant model depends on the block edges and signal level for all blocks. Treating the levels as *nuisance parameters*, and marginalizing them, reduces the full problem into a much more tractable *combinatorial optimization* task—in a nutshell, finding the optimal number of blocks and their edges.

The posterior probability of model M, given data D, is $P(M, \phi, \theta|D)$, where the model parameters have

[1] Of course we sharply distinguish between noise in the sense of random variations inherent in the source and random observational errors.

[2] We adopt the slogan: *the Statistically Significant Structure, the whole Statistically Significant Structure, and nothing but the Statistically Significant Structure.*

been divided into two types: nuisance parameters, denoted by θ, and the others, denoted ϕ. Marginalization of the nuisance parameters is effected simply by carrying out the integral

$$P(M,\phi|D) = \int P(M,\phi,\theta|D)d\theta \ . \qquad (1)$$

Bayes' theorem allows this to be written

$$P(M,\phi|N,V) \propto \int P(N,V|M,\phi,\theta)P(M,\phi,\theta)d\theta \ , \qquad (2)$$

where we have replaced D with the two relevant parameters (N and V, defined below) and eliminated the factor $P(D)$, irrelevant for model comparison since it is independent of the model. We choose the parameters ϕ to be those specifying the edges of the model segments, leaving all others to be treated as nuisance parameters—the most important of which is the parameter representing the constant value of the signal in the block under consideration.

A useful example is the case where the data comprise events, or counts of events, at various locations in the data space, modeled as Bernoulli or Poisson point processes. Marginalizing the event rate parameter characteristically yields a posterior that depends on two *sufficient statistics*: N, the number of events in the block, and V, the size of the block. For event data the posterior of the block model (abbreviated B), marginalized and conditional on the data, is

$$P(B|N,V) = \frac{\Gamma(N+1)\Gamma(V-N+1)}{\Gamma(V+2)} \qquad (3)$$

This quantity can be thought of as the weight[3] which the data gives to model B. The product over the blocks making up a partition gives its weight relative to other partitions. For binned data

$$P(B|N,V) = \frac{\Gamma(N+1)}{(V+1)^{N+1}} \qquad (4)$$

The reader is referred to [13] for the details of this computation, including a discussion of the prior distribution for the signal strength and the units in which V needs to be expressed, and details of the fitness functions for several data modes. Applications are discussed in [14–16].

Nothing in the derivation of the above fitness functions depends on the dimensionality of the data space. For event data in a space of dimension ν, *e.g.*, all that matters is that the expected number of events in an elementary ν-dimensional volume element is equal to a constant (the Poisson rate) times the volume. Hence Eqs. (3) and (4) are valid in any dimension.

[3]The ratio of such weights for two models, called the *Bayes factor*, gives the models' relative probabilities.

5. Prior on the Number of Blocks

One parameter not marginalized, namely the number of blocks, N_b, has a special status, since its value determines the number of other parameters in the complete model. That the value of N_b is automatically found in the optimization is one of the advantages of the dynamic programming algorithm over most cluster analysis methods, in which finding the number of clusters is a vexing problem. One approach is to introduce a term in the fitness function that applies a larger penalty to more complex models. There are various justifications for particular forms of such a penalty term, *e.g.* based on the Minimum Description Length principle [12]. In the Bayesian formalism, there is no need to introduce a penalty term *ad hoc*, since the marginalization of the nuisance parameters yields a built-in effective complexity penalty—sometimes described as the *Occam factor*. But we do need to prescribe a prior distribution for this parameter.

We have adopted a *geometric distribution* for this prior:

$$P(n_b) = C\gamma^{-n_b} \qquad (5)$$

(for $n_b \geq 0$) advocated in [2]. This form yields the following contribution to the log-posterior (ignoring an overall constant):

$$log[P(n_b)] = -n_b \ log(\gamma) \ . \qquad (6)$$

Note that Eq.(6), since it corresponds to subtracting the constant $log(\gamma)$ from the fitness function for each block, trivially maintains block additivity of the fitness function. We are investigating how the strategy of the algorithm might be modified to allow the use of other functional forms for this prior.

6. The Optimization Algorithm

The next step is to optimize the model by maximizing a measure of its goodness of fit over all possible partitions. In [7] we presented a way to find the global optimum of any block-additive fitness function, over all 2^N possible partitions of a 1D interval containing N data points, in time $O(N^2)$. This section is a brief description of this inductive 1D algorithm and its extension to higher dimensions.

Suppose we have the optimal partition of the first n data points, and the corresponding optimal fitness value. Now add one data point, and seek the optimal partition of the first $n+1$ points. Let j be an arbitrary index between 1 and $n+1$, and consider the partition consisting of two parts: (a) the optimal partition of the the first $j-1$ data points, followed by (b) a single block from j to $n+1$. Part (a) and its fitness were found and saved earlier, at iteration number $j-1$, and the fitness of (b) is easily computed. A simple

argument shows that the optimal partition for $n+1$ data cells corresponds to the value of j that maximizes the combined fitness of (a) and (b).

This algorithm can be extended to a data space of any dimension. We continue to take partitions to be sets of blocks containing data cells (*e.g.* Voronoi cells defined by the data points), but relax the constraint that blocks be simply connected.[4] If an optimal block turns out to be not simply connected, it is straightforward to identify its simply connected parts. Relaxing the connectedness constraint has the effect that a few isolated data cells may be assigned to the wrong block. For example a data point with unusually close (far) nearest neighbors, due to a rare statistical fluctuation, may be assigned to a higher (lower) density block than the one that it actually belongs to. Clearly the locations of the data cells are now irrelevant to the optimization. This permits us to arrange the cells in a 1D array so that the algorithm described above can be used. Ordering by cell density—$\rho(c)$ is the number of events in cell c (usually 1) divided by the volume of c—is reasonable, because the piecewise constant model obviously tries to collect together cells with similar densities. This idea is made rigorous by the *intermediate density property*: given three cells c_1, c_2, c_3 ordered by density, $\rho(c_1) \leq \rho(c_2) \leq \rho(c_3)$, if both c_1 and c_3 are in block B_k, an element of an optimal partition, then c_2 is also in B_k. We have proven that this result follows from a certain convexity property possessed by many fitness functions.

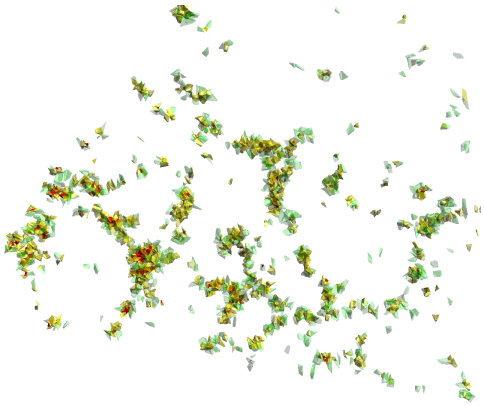

Figure 1: 3D Bayesian Block representation of a section of data from the Sloan Digital Sky Survey. A relatively high density threshold has been set,' revealing the skeleton of the distribution.

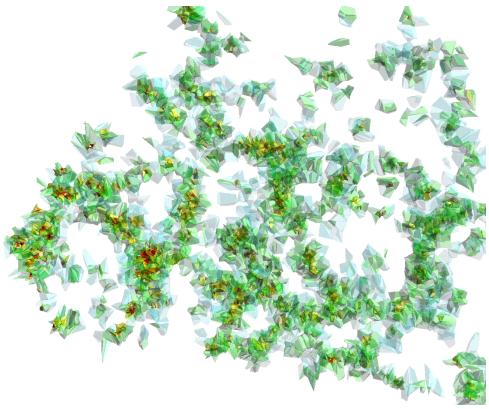

Figure 2: As in Figure 1, but with a lower density threshold, revealing the degree to which these large scale structures are interconnected.

7. An Example and Other Work

We have applied this methodology to a variety of density estimation problems in 1D (mainly time series and the construction of adaptive histograms), 2D (e.g., data from sky surveys), 3D (e.g. data from redshift surveys) and higher dimensions. Space does not permit more than brief mention of one example. Figure 1 shows the Bayesian block analysis of a data set consisting of three dimensional rectangular coordinates of the galaxies with measured redshifts in the first data release from the Sloan Digital Sky Survey. These data are confined to a relatively narrow range of declination, and thus represent a fairly thin slice, here shown in a view perpendicular to the slice. We are developing visualization methods for this block representation, to provide an intuitive picture of the galaxy distribution, free of assumptions about the existence of "clusters" with various symmetry properties. Also, directly from the block representation or by transforming it, one can compute a large variety of derivative statistical quantities describing the 3D galaxy distribution and its topology—correlation and clustering statistics, biasing, genus and genus-related statistics, Minkowski functionals, etc. ([3, 4])

An early version of Bayesian Blocks, based on the greedy algorithm, is in the Astrophysics Source Code Library at `http://ascl.net/block.html` Michael Nowak has developed S code implementing Bayesian blocks in 1D for the S-lang/ISIS Timing Analysis Routines (SITAR) home page `http://space.mit.edu/CXC/analysis/SITAR/` for the Chandra Science Center at MIT. A number of observers have used this approach to study time series data [1, 6, 9, 17, 18].

[4] A set A is *simply connected* if for any partition into two subsets, A_1 and A_2 ($A_1 \cup A_2 = A; A_1 \cap A_2 = \emptyset$), at least one cell in A_1 is adjacent to at least one cell in A_2. For Voronoi cells either of two notions of adjacency can be used: sharing at least one vertex, or sharing at least one face.

Acknowledgments

I am especially grateful to Tom Loredo for many extremely helpful comments on content and presentation. The Applied Information Sciences Research and Intelligent Systems Programs of NASA have supported this work.

References

[1] Bauer, F., and Brandt, W. (2003), "Chandra and HST Confirmation of the Luminous and Variable X-ray Source IC 10 X-1 as a Possible Wolf-Rayet, Black-Hole Binary," Ap. J. Lett., in press, http://arxiv.org/abs/astro-ph/0310039

[2] Coram, M. A., *Nonparametric Bayesian Classification*, Ph.D. thesis, Department of Statistics, Stanford University, 2002.

[3] Park, C., Gott, J., and Choi, Y., "Topology of the Galaxy Distribution in the Hubble Deep Fields," (2001), **Ap. J.**, 553, 33

[4] Hikage, C. et al, "Minkowski Functionals of SDSS Galaxies I: Analysis of Excursion Sets," (2003), **P.A.S.J.**, 55, 911

[5] Ebeling, H., and Wiedenmann, G. (1993), "Detecting structure in two dimensions combining Voronoi tessellation and percolation," *Physical Review* E, Volume 47, pp.704-710.

[6] Hambaryan, V., R. Neuhaeuser, R., and Stelzer, B. (1999), "Bayesian flare event detection: ROSAT X-ray observations of the UV Cetus type star G 131-026," Astron. Astrophys., **345**, pp. 121-126

[7] Jackson, B., Scargle, J., Barnes, D., Arabhi, S., Alt, A., Gioumousis, p., Gwin, E., Sangtrakulcharoen, P., Tan, L., and Tao Tsai, T. (2003) "An Algorithm for Optimal Partitioning of Data on an Interval," submitted http://front.math.ucdavis.edu/math.NA/0309285

[8] Kim, R., Kepner, J., Postman, M., Strauss, M., Bahcall, N., Gunn, J., Lupton, R., Annis, J., Nichol, R., Castander, F., Brinkmann, J., Brunner, R., Connolly, A., Csabai, I., Hindsley, R., Ivezic, Z., Vogeley, M., and York, D. (2002), "Detecting Clusters of Galaxies in the Sloan Digital Sky Survey. I. Monte Carlo Comparison of Cluster Detection Algorithms," Astronomical Journal, Vol. 123, pp. 20-36.

[9] Kim, D.-W., Cameron, R. A., Drake, J. J., Evans, N. R., Freeman, P., Gaetz, T. J., Ghosh, H., Green, P. J., Harnden, F. R., Jr., Karovska, M., Kashyap, V., Maksym, P. W., Ratzlaff, P. W., Schlegel, E. M., Silverman, J. D., Tananbaum, H. D., Vikhlinin, A. A., and Wilkes, B. J. (2003) Chandra Multi-wavelength Project (ChaMP). I. First X-ray Source Catalog, in preparation, http://arxiv.org/abs/astro-ph/0308492.

[10] Okabe, A., Boots, B., Sugihara, K., and Chiu, S. N. (2000), *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, John Wiley and Sons, Ltd., New York, Second Edition

[11] Ramella, M., Boschin, W., Fadda, D., and Nonino, M. (2001), "Finding galaxy clusters using Voronoi tessellations," Astronomy and Astrophysics, v.368, p.776-786.

[12] Rissanen, J., 1989, *Stochastic Complexity and Statistical Inquiry*, Singapore: World Scientific.

[13] Scargle, J., 1998, "Studies in Astronomical Time Series Analysis. V. Bayesian Blocks, A New Method to Analyze Structure in Photon Counting Data", *Astrophysical Journal*, **504**, p. 405-418, Paper V. http://xxx.lanl.gov/abs/astro-ph/9711233

[14] Scargle, J. D., (2001), Bayesian Blocks: Divide and Conquer, MCMC, and Cell Coalescence Approaches, in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering,* 19th International Workshop, Boise, Idaho, 2-5 August, 1999. Eds. Josh Rychert, Gary Erickson and Ray Smith, AIP Conference Proceedings, Vol. 567, p. 245-256.

[15] Scargle, J. D., (2001a), "Bayesian Estimation of Time Series Lags and Structure," Contribution to **Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MAXENT 2001)**, held at Johns Hopkins University, Baltimore, MD USA on August 4-9, 2001.

[16] Scargle, J. D., (2001), "Bayesian Blocks in Two or More Dimensions: Image Segmentation and Cluster Analysis," Contribution to **Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MAXENT 2001)**, held at Johns Hopkins University, Baltimore, MD USA on August 4-9, 2001.

[17] Wheatland, M., Sturrock, P., and McTiernan, J. (1998), Ap. J., **509**, p. 448-455.

[18] Wheatland, M. (2000), Ap. J., 536 , L 109-112, 2000, "The Origin of the Solar Flare Waiting-Time Distribution."